Disease Control Priorities Project

Working Paper No. 43
September 2005

Comparing Quality in Disparate Settings Using Vignettes to Control for Case-Mix Variation

John W. Peabody, MD PhD
Institute for Global Health
University of California, San Francisco
Peabody@psg.ucsf.edu

Anli Liu
University of California, San Francisco
Anli.liu@ucsf.edu

Lily Alisse
Institute for Global Health
University of California, San Francisco
lpanyaco@yahoo.com

Jesus Aiying Sarol, PhD
Manila, Philippines
jsarol@nwave.net

Jane Yunjing Ren, MD
Peking University Medical College
Bejing, China
Janeyjren@hotmail.com

Jorge Munoz, PhD
Escuela Superior de Economicas y Negocios
San Salvador, El Savador
Jorge@lpsfa.com

Carlos Carillo, MD
Instituto Nacional de Salud Publica
Cuernavaca, Mexico
cecarril@insp3.insp.mx

Naveet Wig, MD
All-India Institute of Medical Sciences
New Delhi, India
naveetwig@vsnl.com

Stella Alabastro Quimbo, PhD
School of Economics
University of the Philippines
Stella_quimbo@yahoo.com

Corresponding Author:
John W. Peabody, MD PhD
74 New Montgomery, Suite 508
San Francisco, CA 94105
Tel: (415) 597-8202
Fax: (415) 597-8299

Peabody@psg.ucsf.edu

The Disease Control Priorities Project is a joint effort of The World Bank, the Fogarty International Center of the National Institutes of Health, the Bill & Melinda Gates Foundation, and the World Health Organization.

ACKNOWLEDGEMENTS

This working paper is the product of an international collaboration. We are indebted to the following individuals for their help and support. We would like to thank Dan Bertenthal for conducting the preliminary analysis and to Lin Zhang for completing the analysis. In China, we would like to thank Dr. Zhang Kong Lai; in El Salvador, Dr. Jaime Avila; in India, Dr. Narendra Arora, Dr. Neerja Bhatla, Dr. Rakesh Lodha, Yogesh Singh, Alok Kumar Gupta, Chander Mohan Mittal, and Saurabh Jha; in Mexico, Dr. Stefano Bertozzi, Dr. Ofelia Poblano Verastegui, and Juan Pablo Gutierrez; in the Philippines, Dr. Mario Taguiwalo, and Oliver Asis. These individuals helped us to navigate the respective health systems of each country, obtained the ethical clearance, and collected the data. We would also like to thank Sarah Weston and Cindy Muzio for their assistance preparing the manuscript.

INTRODUCTION

Two principles ought to guide the discussion about improving the quality of care. First, the quality of clinical practice should be measured using evidenced-based criteria grounded in well-articulated theory and substantiated by valid studies. Second, better quality should, in turn, lead to improved health outcomes for patients. It has been surprisingly difficult to show a direct relationship between better-quality care and better health outcomes such as lowered morbidity and mortality rates, particularly at the population level (Donabedian 1988; Peabody, Rahman et al. 1999; Salmon, Heavens et al. 2003). A few recent studies, however, suggest that delivering high-quality care (that is, consistently complying with a standard set of evidence-based health practices) leads to improved survival rates (Skinner, Fisher et al. 2001; Wennberg, Fisher et al. 2002). In the Netherlands, for example, the implementation of clinical guidelines in general practices has been linked to improved outcomes in patients with asthma and chronic obstructive pulmonary disease (Jans, Schellevis et al. 2000; Jans, Schellevis et al. 2001).

In studies that compare clinical practice to evidence-based criteria, researchers have found that even in developed countries, high-quality care is inconsistently provided to large segments of the population (McGlynn, et al 2003; Nagaya, Fetters et al. 2002; Wennberg, Fisher et al. 2002). A study in the United States found that physician compliance with evidence-based guidelines exceeded the 80% level in only 8 of 306 regions; compliance in 10 regions stood at less than 10% (Wennberg, Fisher et al. 2002). And a comprehensive literature search that evaluated the quality of health care in large or diverse populations within the United States showed wide variation among hospitals and geographic locations (Schuster, McGlynn et al. 1998). In the treatment of acute myocardial infarction (AMI), for example, not only was there substantial geographic variation, but therapies of proven benefit in AMI (medication, smoking cessation advice, and reperfusion) were consistently underused (O'Connor, Quinton et al. 1999). Quality of care has also been shown to vary by other factors, including provider characteristics, reimbursement scheme, and system of

care (Donohoe 1998; Harrold, Field et al. 1999; Diette, Skinner et al. 2001; Seddon, Ayanian et al. 2001; Kerr, Gerzoff et al. 2004).

These variations are often independent of access to care or the costs of care. For instance, in a direct geographical comparison, greater Medicare spending in the United States was not associated with longer survival rates. Importantly, however, the same study showed that indicators of effective practice in a given geographic region (such as vaccination for pneumococcal pneumonia, colon cancer screening, and eye examinations for diabetics) did result in better health outcomes, so that for every 10% increase in the index of effective practices, survival rates improved by 0.2% (Skinner, Fisher et al. 2001).

There is emerging evidence that similar variation exists in developing countries as well (Loevinsohn, Guerrero et al. 1995). Variation appears to exist across facilities, among providers, and between specialists and nonspecialists in developing countries (Walker, Ashley et al. 1988; Beracochea E 1995; Peabody, Gertler et al. 1998; Nolan, Angos et al. 2001; Weinberg 2001; Dumont, De Bernis et al. 2002). For example, in an evaluation of district and teaching hospitals in seven developing nations (Bangladesh, Dominican Republic, Ethiopia, Indonesia, Philippines, Tanzania, and Uganda), triage services were rated as good or excellent in only half the teaching hospitals and less that one-fourth of the district hospitals (Nolan, Angos et al. 2001). Even though the data are limited, it is striking how much variation there appears to be in the quality of care within individual countries. This observation contradicts established notions that care is better in some countries compared to others. With so much variability *within* countries, it seems unlikely that variation *among* countries is very meaningful.

We thus hypothesized that, for the same clinical conditions, quality of care would vary widely within different countries. To test this hypothesis, we measured the quality of clinical practice in five countries having different systems of organizing and financing care—China, India, the Philippines, El Salvador, and Mexico. We measured quality for three common clinical conditions that are prevalent in all study sites. To measure quality, we used clinical vignettes—written case scenarios administered to doctors—because they are a validated measurement tool

that controls for all case-mix variation and thus accounts for variations in individual health status that would otherwise confound a cross-national survey (Peabody, Luck et al 2004b). We were therefore able to obtain an identical measure of quality across all countries. Doctors for the study were randomly selected from separate rosters generated at four different types of facilities—tertiary care hospitals, district hospitals, public outpatient clinics, and private clinics. We developed overall quality scores that measured both the average and the range of clinical practice in the contexts of physician characteristics, physical setting, facility type, and country. We also determined variation in specific skills by medical condition and physician income.

Defining Quality

While "high quality care" is a desirable attribute of clinical practice, it is often unclear exactly what that term means (De Geyndt, 1995). Quality does indeed encompass a broad range of issues and is used in many different ways. Thus, precision is warranted in discussions about measuring and improving quality. The Institute of Medicine defines *quality* as "the degree to which health service for individuals and populations increase the likelihood of desired outcomes and are consistent with professional knowledge." (IOM 2001) Quality medical care, by extension, involves the management of a patient's health benefits and risks through the provision of technologically sound care (Donabedian 1980). Given a patient with a particular health profile, a medical intervention is considered *appropriate* if its expected benefits (e.g., increased life expectancy, reduced pain, improved function) outweigh its expected risks (e.g., pain, morbidity, mortality, cost) by a substantial margin. A subset of appropriate care is *necessary* care—which provides a significant benefit and which it would be considered improper to withhold. Conversely, inappropriate or poor-quality care can mean providing too much care (through unnecessary tests and medications, with associated risks and side effects), too little care (through not providing an indicated diagnostic test, pain relief, or life-saving procedure), or the wrong care (improper treatment techniques)(Shuster, McGlynn et al. 1998).

The Elements of Quality

Structure. The structural elements of quality refer to the stable characteristics of providers, their tools and resources, and the organization needed to provide care (Donabedian 1980). Structural inputs are the most commonly measured elements of quality in developing countries. This is because, practically speaking, structure is often the easiest way to evaluate quality. Structural assessments are also easy to understand and thus a useful descriptor—especially when many facilities lack basic equipment, supplies, or staff. (Donabedian 1980; Peabody, Rahman et al. 1994).

At best, however, structural inputs are a blunt means of measuring quality. Because they are proximal to process, structural elements have a lesser impact on outcome. The presence of structure, i.e., the availability of doctors, infrastructure, and health insurance, can only be viewed as an *intent* to provide health care services. While the presence of structure may increase the probability of a good outcome, this relationship must be firmly established before greater investments in health care can be justified as a way to improve health outcomes (Brown 1988). The weak causal relationship between structure and outcome explains why new physical resources rarely improve the health of the population (Berggren, Ewbank et al. 1981; Peabody, Gertler et al. 1998).

Although closely aligned with structure, access to care (the "broad set of concerns that center on the degree to which individuals and groups are able to obtain needed services from the medical care system" (IOM, 1993, pg. 4)) should be viewed as a separate but important tool for measuring quality. A variety of research has demonstrated that improved access can have a positive effect on health outcomes (Barker 1983; McCormick 1985; Krieger 1989; Boyce 1991). Using applied estimation techniques to evaluate the presence of low-cost health services and midwives, one analysis determined that community factors encouraging utilization had the greatest impact on the probability of survival of 5-year-old children in the rural Philippines (Abejo 1987). In another study that looked at the relationship between the quality of child health care services, access to care, and the survival and health outcomes in children in Ghana, the provision of child

services had a significant, positive impact on the survival of children(Lavy, Strauss et al. 1996). Extrapolation of the results showed that having child services available for one more hour a week (~ 15% increase) would increase median survival duration by almost 1%. But greater access alone will not improve health. For example, no clear relationship exists between the number of prenatal visits and specific outcomes such as birth weight (Blondel, 1985). And access to poor-quality care may not only fail to improve health, it may also cause health to be worse (Druss, Bradford et al. 2001; Goulding 2004). Donabedian offers a feasible solution to the dilemma by choosing not to include access per se into a working definition of the quality of care (Donabedian 1980).

Process. The quality of the process of care can be measured by the degree to which physicians comply with predefined indicators of effective care (Shuster, McGlynn et al. 1998; Wennberg, Fisher et al. 2002). The process of care, or "the set of activities that take place between and within practitioners and patients," occurs frequently— every time a patient and a provider come together. Thus, process is an easier measure to accumulate than an adverse health event. Process is also relatively easy to observe in a clinical setting, and specific changes in clinical practice can be evaluated for changes in outcome (Peabody 1995). The process of care is often further divided into technical care and interpersonal/artistic care.

The Technical Component: The application of specific technical skills transforms inputs into the actual diagnosis or clinical intervention. Although the correct application of skills should be based on scientific investigation, the provider can never be certain that an intervention will alter the course of an illness. Nevertheless, measuring the process of care is most commonly focused on measuring the application of these specific scientific skills (Bryce, Toole et al. 1992).

The Artistic Component: By contrast, the interpersonal or artistic elements of clinical care are more difficult to measure. Artistic elements include warmth, confidence, and judgment in the face of technical uncertainty. While this aspect of care is clearly important, it is more qualitative in nature and becomes even more challenging to assess when elements of culture are included (Collins, Clark et al. 2002; Silverman, Terry et al. 2002).

Structural and process measures are thus potentially useful markers of better health outcomes. In a study comparing the contribution of the process and structural elements of antenatal care on birth weight outcomes, process was found to matter more than structure. More specifically, better clinical examinations resulted in higher birth weights across the population than the structural qualities of the facility (Peabody, Gertler et al. 1998). Structural inputs, such as the availability of basic medical equipment, affected outcomes only when they were necessary to providing good care. On the other hand, structural inputs with more distal relationships to outcomes did not seem to have an effect on birth weight, nor did they explain and general variations in process (Peabody, Gertler et al. 1998). In a similar study, maternal health outcomes were significantly and positively associated with the degree of training of the health attendants/midwives (Dumont, De Bernis et al. 2002). Such a positive relationship was attributed to the fact that the midwives in health facilities detected more obstetric complications, thus leading to immediate care and a decrease in the case fatality rate.

Outcome. Outcome is defined as the change in a patient's current and future health that can be attributed to the quality of care. This definition of outcome renders a useful, though simplistic, view of outcome measures and how outcome can be affected: Improving quality of care avoids poor health outcomes. Nevertheless, although changes in outcome are clearly a direct measure of the quality of care, they are often not a good way to measure quality. This is because outcomes tend to be multi-factorial, and it is therefore difficult to separate out the effect of quality of care from other intervening factors (Peabody, Tozija et al, 2004)**.** Factors ranging from an individual's genetic makeup to community-level variables all have an effect on health outcomes. The low frequency of adverse events (e.g., death) numerically limits the extent to which outcomes can be used. Additionally, improvements in quality may be overlooked if only such events as mortality are measured, as in the case of prenatal care (Peabody 1995; McGlynn, Asch et al. 2003).

Measuring Quality of Care

Measuring the level of quality in rich and poor countries alike has been difficult (McGlynn, Asch et al. 2003). Not one of the established measurement methods is without drawbacks (Badger, deGruy et al. 1995; McDonald, Overhage et al. 1997; Peabody, Luck et al. 2000). It is probably most useful to categorize them on the basis of their limitations and strengths, which indicate where and how they can best be used. Chart abstraction, direct observation, recorded visits, administrative data, standardized patients and vignettes are the most commonly used methods. A description of these methods' costs, biases, data collection, ethical issues, and ability to account for case-mix variation are summarized in Table 1 and detailed below.

Chart Abstraction. Quality measurement has historically relied on medical chart reviews (McDonald, Overhage et al. 1997). Charts are typically available after each encounter and can be abstracted by a trained professional. Their fundamental limitation is that they do not accurately reflect actual practice. In a study comparing methods of measuring quality of care, chart abstraction was found to underestimate the quality of care for common outpatient general medical conditions, especially as it compared to standardized patient reports(Luck, Peabody et al. 2000). Medical charts were also often found to be incomplete and not reflective of all the events that transpired during the clinical visit (Dresselhaus, Luck et al. 2002). Other problems associated with using medical records in different systems and countries are multiplied by differing record-keeping practices. They vary not only from health care setting to health care setting but also from country to country (Peabody, Tozija et al. 2004). Precise and reliable chart abstraction can also be a costly process because it involves ensuring that the medical abstractor is medically sophisticated, is systematically trained, and meets regularly with researchers to review criteria.

Direct Observation and Recorded Visits. Direct observation or recorded visits, where the actual interaction between provider and patient are recorded or observed by a trained abstractor, are feasible alternatives for measuring quality of clinical practice. The major shortcomings associated with these similar methods are the costs, variation between observers and the introduction of the social desirability bias—observed physicians improve their performance in

ways that do not reflect their actual practice. In addition, ethical considerations dictate that the patient and the provider need to be informed of the planned observation. Either or both the patient and the provider might find the situation uncomfortable and act differently from usual.

Administrative Data. The use of administrative data to measure the process of care involves extracting information from existing patient databases. These databases are often generated as part of cost accounting and are becoming more and more available in poorer countries. This method also suffers from important limitations. For example, while information on patient diagnosis and treatment is generally included, notes on history taking or physical examination are not. The main problem, however, with using administrative data is inaccurate diagnoses because administrative forms are incorrectly or impartially completed. Finally, information on patients' comorbid conditions is generally scarce and thus does not account for case mix (Peabody, Luck et al. 2004 a).

Standardized Patients. Standardized patients (SPs) are considered the gold standard for measuring quality of care (Rethans and van Boven 1987; Colliver, Vu et al. 1993; Badger, deGruy et al. 1995; Colliver and Swartz 1997; DeChamplain, Margolis, et al. 1997; Luck and Peabody 2002). SPs are trained actors who simulate a medical illness and present themselves unannounced into a clinical setting. At the conclusion of the visit, they report on the technical elements of the process of care. SPs have been shown to provide accurate and consistent measures of physician performance and can capture variation in clinical practice. They also reproducibly show how individual physician practices vary over time (Colliver, Vu et al. 1993; Swartz and Colliver 1996; Carney and Ward 1998; Glassman, Luck et al. 2000; Luck J and Peabody JW, 2002). SPs, however, are prohibitively expensive and only useful for a limited number of adult outpatient conditions; thus, they are not viable for routine use in evaluating quality of care (Colliver and Swartz 1997; Glassman, Luck et al. 2000).

Clinical Vignettes. Clinical vignettes are well-suited for measuring the quality of care, particularly in population studies. Clinical vignettes are simulated cases administered to physicians—on paper or via computer—that measure the technical elements of the process of care.

(Palmer, Louis et al. 1985; Goldman 1992; Peabody, Gertler et al. 1998; Peabody and Luck 1998; Peabody, Luck et al. 2000; Peabody, Luck et al. 2004b; Peabody, Tozija et al. 2004). Clinical vignettes overcome three specific problems apparent in other methods of measuring quality of care: case-mix adjustment for the range of clinical severity and core sociodemographic factors, disparate methods of medical record-keeping, and undue emphasis on inputs or the structural elements of care (Salem--Schatz, et al 1994; Dresselhaus, Peabody et al. 2000). Past experience in administering vignettes has shown that doctors are generally cooperative and refusal rates are low. These advantages make them particularly useful for cross-system comparison and for evaluation of care in developing countries. Their main limitations are that they are limited to a single visit and place an additional modest burden on a doctor's time (Fihn 2000).

Quality Measurement at the Population Level

Thus far, we have focused on the physician-patient dyad. To evaluate variation in the quality of care among and within countries, however, we are also interested in quality measurement that goes beyond personal doctor-patient services to the population level. This is because, at the broadest level, the purpose of policy is to implement behavioral change across *groups* of providers. We also want to avoid a situation whereby we improve the health status of one person in a group of patients by inadvertently limiting the care available to another. Finally, successful improvements in quality are should lead to better health at the population level and not just among a few individuals.

Measuring health outcomes in aggregate is a long-established practice—for example, with infant mortality rates and birth weight. The technical element of process among a group of providers can be similarly measured in aggregate by assessing provider compliance with evidence-based, expert criteria. Disease-specific criteria can be selected on the basis that their correct application is known to lead to better health outcomes (e.g., aspirin in the post–myocardial infarction patient). Then, for any given case, the percentage of evidence-based criteria fulfilled can

be used to determine a provider's quality score. At the population level this can be done for a group of providers caring for a population of patients.

Such an approach assumes that the measurement method adjusts for case mix. It also argues for an assessment of clinical conditions that are prevalent (i.e., have a large burden of disease) and those for which effective treatment is available. Finally, cost and feasibility must be considered when defining the intervention criteria that are used to produce better health.

The quality of clinical practice at the community level must be represented by aggregated data on the process of care across a group of providers. Thus, any assessment should take place at the major centers of care within the community—the tertiary care hospital, the district care hospital, and public and private clinics. While each health system has its distinctive features, this basic stratification is found throughout much of the world. Generally, patients present themselves to the generalist at the private or public clinic. More advanced or chronic cases or conditions that require inpatient support are referred to generalists or specialists at the district care hospital. Finally, the most difficult cases are referred to specialists working at the tertiary care centers. Obviously, patients often jump to district or testing facilities so that, practically speaking, care for common conditions must be provided at all levels. Nevertheless, each successive level not only becomes more specialized in its care but draws from greater structural inputs, serves a larger geographical catchment area, and handles a larger and more diverse patient population.

Although aggregated data on the quality of care within a community are needed as a foundation for policy reform, such data are inherently problematic. Often, summary data are not readily available. Moreover, they can never completely represent the average health care system of the entire community (whether it be a small rural community or a country). When comparing quality between and among countries, it is neither feasible nor informative to try to measure a representative sample of physicians in the entire country. But because the sample frame is limited, variation, will, by definition, be underestimated.

Hypothesis

Our study looked at the quality of care in five different countries: China, India, Philippines, El Salvador, and Mexico. We hypothesized the following:

- A wide variation in process of care exists within each country, as measured by vignettes, regardless of condition or clinical skill.
- The variation within countries is greater than the difference between countries.
- Disparities in quality of care exist between public and private providers as well as between district and tertiary facilities.
- The variation in quality of clinical care is associated with specific physician characteristics.
- Because quality of care is difficult to observe, quality of care in not reflected in physician income.

METHODS

Study Design

We conducted a prospectively designed evaluation of quality of care among randomly selected physicians. The physicians were located in four distinct health settings: tertiary care hospital, district level hospitals, and public and private outpatient clinics. Five countries participated in the study: China, India, Philippines, El Salvador, and Mexico. Data were collected between June and August 2003. Within each country, we specified the four settings by first choosing the public tertiary care hospital with a reputation for providing the best medical care in that country. The reportedly "best" public tertiary hospitals were located in the largest or second-largest urban centers of each country.

Sample

We used a stratified sample frame that was based upon the selection of the leading public tertiary care facility in each country. These facilities accepted referrals from other urban district level hospitals. To determine which district hospital to sample, we created a roster of referring district hospitals situated within a 60-mile radius (100km) of the referral hospital. Hospitals were included on the list if they provided inpatient care, were within the defined radius, were publicly financed, and regularly referred specialty and advanced care cases to the identified leading public

facility. We used similar criteria to select public and private clinics. The clinics had to provide outpatient care, be located within 60 miles (100 km) of the referral and district hospital, and regularly refer their patients to the district or (in some cases) to the leading referral hospital. To generate the roster of public and private clinics, we used a snowball sampling technique and obtained lists of clinics or practitioners by asking regional health authorities, looking at official rosters of facilities, searching the phone book, and contacting medical societies. Once placed on the roster, both public and private clinics were contacted at random and asked if their doctors would participate in the study. The top tertiary hospitals, used as the reference facility in the sample frame, are listed by country in Table 2.

Inclusion Criteria for Doctors

We sought to be inclusive and defined broad eligibility criteria for participation in the study. Non-medical doctors were excluded from the study because of the wide variation in licensing and practice that exists between countries. Doctors were eligible if they

- had a license to practice medicine
- had graduated from a nationally accredited medical school
- provided relevant (to the cases) specialist or primary care
- predominantly practiced in one of the four specified settings
- voluntarily agreed to participate.

Our samples were drawn as follows:

Tertiary Care Doctors. From the leading public hospital in each country, we obtained a complete list of doctor specialists in obstetrics, pediatrics, and internal medicine. Eight doctors from each specialty were then randomly selected from these lists and asked to participate by completing a vignette related to their area of specialty.

District Hospital Doctors. For each district-level facility, we generated a roster of generalists caring for patients to select eight doctors. If the standard of care for generalists did not include obstetrical care, we additionally created a roster of obstetricians working at the district

level to select eight additional obstetricians. The eight generalists were asked to complete vignettes for all three cases (a doctor who did not care for pregnant women was evaluated for two cases and an obstetrician was evaluated for the prenatal case).

Public and Private Clinic Doctors. At the clinic level, we generated sample rosters using a snowball technique. This yielded a large number of public and private generalist doctors. From these aggregated lists, eight doctors were randomly selected. If a doctor was not available when we visited the clinics or when we set the time to administer the vignette, we substituted the next name on the randomly generated roster. As before, if a provider in the area did not provide obstetrical care, a specialist providing public or private outpatient obstetric care was selected and asked to complete the prenatal case evaluation.

A total of 480 vignettes were administered to the 300 doctors who agreed to participate in the study. The refusal rate was 7%.

Recruitment

Doctors were solicited in person and by telephone. Participation in the study was strictly voluntary and completely confidential. Participants received an information sheet outlining briefly the purpose and the background of the study, procedures, all risks and discomforts, and benefits and costs involved in the physicians' participation. The participant information sheet included the principal investigator's contact information and a local phone number for the country-specific site coordinator, when applicable. The overall study was approved by the Institutional Review Board (IRB) at the University of California, San Francisco. Local IRB authorization was sought at each site and obtained on a case-by-case basis in coordination with local authorities. Before participating in the study, doctors were told they could withdraw at any time. No names were linked with responses.

Conditions and Diseases

We used three clinical cases in the study: prenatal care, diarrhea, and tuberculosis. Selection of these cases was based on the following criteria:

- They represented common outpatient conditions or diseases in the developing world.
- They had an associated high burden of disease.
- In each case, higher quality of care (better process) had been demonstrated to lead to improvements in health outcomes.
- Appropriate technology was available in all sites.

Site Characteristics

Because we recognized that it was not practical or feasible to collect a complete sample or even a random sample of countries, the countries were purposely selected to represent a broad spectrum of health care in the developing world. Specifically, we were interested in countries that exhibited a wide variation in overall health status, health care systems, and basic country statistics such as population size. Table 3 lists some summary health statistics as they relate to participating countries and the conditions or diseases that we used as the subject of our vignettes.

We also sought countries that represented different geographic regions of the world with varying demographics. The countries ranged in population size from slightly over 6 million in El Salvador to over a billion in both China and India. As of 2002, Mexico had attained the highest life expectancy, at 71.7 years for males and 77.0 years for females. On the other end of the spectrum, life expectancy in India was just 60.1 years for males and 62.0 years for females. Mexico and China's relatively low child mortality rates contrasted sharply with India's high rates of 87 deaths per 1000 males and 95 deaths per 1000 females (WHO 2003). Maternal mortality rates per 100,000 live births were notably elevated in China, the Philippines, and India. Tuberculosis, currently the leading cause of death in India for the 15-45 age group, claims close to 400,000 deaths per year (World Bank Group 2004). The Philippines, however, had the highest estimated prevalence of tuberculosis cases per 100,000 people, at 540 (WHO 2003).

Organization of Care. Although the five countries exhibit different systems of organization, patient referral, and distribution of care (summarized briefly below), they all have an extensive system of publicly provided (and financed) care. The extent to which the health care system is publicly provided is determined by policy and is reflected in the costs and availability of

services. In all countries, there is also a large component of privately provided services, and these continue to exert pressure toward market-based health systems. China, for example, has allowed much of its economy to move toward a market based health system. A large majority of its hospitals, however, remain government run, and clinical care is publicly provided but financed both publicly and by out-of-pocket (private) payments. In New Delhi, India, and Manila, in the Philippines, there is a high density of both public and private hospitals as well as private and public clinics; a majority of facilities in these two countries are privately run. While patients may receive free health care from the public tertiary care facility and public clinics, they usually pay for medicines out-of-pocket. They may also elect to receive private health care on a fee-for-service basis within the private system. The extent of medical coverage, including medication, varies but does not cover all costs. El Salvador and Mexico have a national social security system that covers all workers and a national public health system that in principle provides coverage to all citizens. Out-of-pocket payments tend to be high, particularly for medication.

Patient Referral. Over the past 20 years, the international trend has been toward increasing specialization, particularly in hospital care (WHO 2000). As is the case with many of the countries included in our study, the hospital care system in and around the large urban centers, such as San Salvador, Beijing, and New Delhi, is already very specialized. In New Delhi, for example, a doctor in a district level hospital may be licensed as a general care physician, but may practice mainly in pediatric care, obstetric care, or respiratory illness.

Provision of Obstetric Care. The provision of care is strongly affected by cultural and gender preferences. In El Salvador and India for example, obstetric care is separate from the duties of the general practitioner. Public clinics in Beijing, China, do not provide obstetric care in the same location as other services. In some parts of India, where gender relations remain fairly conservative, pregnant women overwhelmingly prefer to see female obstetricians. Therefore, not only are most obstetricians female, but male general practitioners will frequently refer pregnant women to a female obstetrician for prenatal care.

Financing. Health care services in all five countries is provided through the combination of public and private measures. China's health care services are primarily publicly funded, although public clinics are typically both public and private. China's doctors are salaried; private practice, although a reality, is seldom acknowledged. Likewise, the Salvadorean system is publicly financed by the Ministry of Public Health (covering 83% of the population) while 17% of the population working in the private sector is covered by the Social Security Institute. The Philippine health care system, by contrast, is mostly privately financed on a fee-for-service basis. If Filipinos elect to receive their care from publicly run facilities, they encounter long waits and few medications. In India, people rely on private spending for health care more than almost anywhere else in the world. Paying for health care in this manner can be problematic: More than 40 percent of Indians need to borrow money or sell assets when they are hospitalized. Furthermore, even when citizens receive free health care at a public hospital, they must pay for medications out-of-pocket (A Vision for India's Health System Conference 2001). Mexico's health system has been characterized as a patchwork of multiple parallel public and private arrangements. This balkanization of health care may explain why half of Mexico's 100 million citizens are uninsured, over half of the annual expenditure on health care is out-of-pocket, and why there is such disparity in health outcomes between the richest and poorest parts of the country. Employed people are caught between an uneven and disjointed system of public-private coverage, while a good portion remain uninsured.

Data Collection

To operationalize the project locally, a local site/project coordinator in each country completed the following tasks:

- Obtained approval from a local institutional research (review) board (IRB) and/or local authorities.
- Developed a roster of doctors at the four sites.
- Began recruitment at the four sites where doctors would complete clinical vignettes.
- Established a schedule for data collection.

To account for the differences in local IRBs, the local site coordinators were charged with the responsibility for identifying and obtaining ethics approval from the appropriate institution. Jurisdiction over research and research structures varied from country to country. For example, the respective departments, ministries or regional/provincial divisions of health typically gave ethics approval in writing, whereas approval from individual hospitals or clinics frequently came were given verbally.

Vignettes

To measure the quality of clinical care, 488 vignettes were given to doctors in the four sites located in five different countries. Eight vignettes either were not completed or were left blank; they were dropped from the analysis, leaving a total of 480 completed vignettes. The vignettes were organized in sections or skill domains designed to recreate the sequence of a typical patient visit: presentation of the patient and his or her medical complaint, history-taking, physical examination, radiological or laboratory tests ordered, diagnosis, and treatment plan. In each domain, doctors were asked open-ended questions—for example, "What information would you obtain when taking the patient's history?" and "What would you look for during the physical examination?" Once physicians had completed a domain, they could not return to a previous domain to revise their answers or use new information given after a domain was completed to change (or improve) their previous answers.

Vignette Validation

The vignettes used in the study were previously validated (Peabody, Luck et al. 2000; Peabody, Luck et al. 2004b) by the use of standardized patients (SPs) as the "gold standard" for measuring clinical practice (Rethans and van Boven 1987; Colliver, Vu et al. 1993; Badger, deGruy et al. 1995; De Champlain, Margolis et al. 1997). The first step of the validation procedure was for standardized patients to be introduced unannounced into clinics (the detection rate was 5% in one study) (Peabody, Luck et al. 2000). Immediately after each physician visit, the SPs completed a checklist, indicating exactly what actions the doctor took during the visit. The checklist generated

an SP score. That score was then compared to the same physician's clinical vignette score, which was calculated according to identical criteria**.** Each SP visit also generated a medical record that was retrieved by the study team. This abstracted chart was scored using the same criteria as the other two, thus generating a third score. The SP checklist, the medical record from the SP visit, and the corresponding vignette completed by the physician were then compared to determine the validity of each measurement method.

The results from two large prospective validation studies showed that vignettes consistently produced scores closer to the gold standard of SPs than did the abstracted medical charts ($p > .05$) (Peabody, Luck et al. 2000; Peabody, Luck et al. 2004b). Analyses confirmed this finding to be robust across sites, case, complexity, and level of training ($p > .05$).

Scoring Criteria

To create the scoring criteria, we conceptualized quality as the comprehensive provision of services for a given clinical case in a manner that leads to better outcomes for individuals and populations (Peabody, Luck et al, 2004 b). Thus, we identified candidate quality criteria for a full range of provider activities that make up the process of outpatient primary care and have been shown in the evidence-based literature on quality of care to lead to better outcomes. This involved describing the complete set of actions that would need to be undertaken by physicians when they saw patients. Thus, single-point measures, such as determining whether an antibiotic was prescribed or whether the patient was screened in the history for a comorbidity, were not the sole determinants of a doctor's quality score. Instead, comprehensive criteria were developed for each of the five domains of care: (1) taking the relevant history; (2) performing the relevant portion of the physical exam; (3) ordering the necessary laboratory or imaging tests; (4) making the correct diagnosis, including the etiology, and (5) prescribing a complete treatment (management) plan.

The quality criteria for each of the three cases were derived from three sources: an evidence-based literature search on the clinical practices that lead to better health outcomes; U.S. and international clinical guidelines; and local expert panels of academic and community

physicians comprising both generalists and specialists. We used recommendations by the expert panels to modify and finalize the master criteria list derived from the literature and guidelines (See Glassman, Luck et al., 2000 for examples.) Items felt by experts to be most critical were assigned a weight of 1.0. Items that experts deemed less important, such as multiple physical examination items related to a single clinical construct, were grouped into categories, which implicitly assigned them lower weights, typically 0.50 or 0.33.

Scoring was done by two trained abstractors. The abstractors were blinded to physician identity. They reviewed each vignette answer sheet and indicated on a scoring form those items the physician had successfully completed. The raw item scores were aggregated into category scores. These weighted scores (an average of 41 categories per case) were than totaled and divided by the total possible score, generating a percentage correct score for each vignette. For further subanalyses, each scoring category was assigned to one of the five domains of the encounter.

Vignette Administration

Before the vignettes were administered in the respective countries, they were translated into Spanish and Chinese and then back-translated to check for consistency. Different pairs of bilingual physicians were used to ensure linguistic and technical accuracy, including local variations in medical terms and care. Prior to scoring, the responses were translated by the same bilingual physicians. Ten percent of the translations were randomly retranslated to ensure accuracy and consistency. Another ten percent were randomly audited and scored a second time to ensure scoring and accuracy. The error rate was 5%, well within usual standards observed from abstraction (Zadnik, Mannis et al. 1998; Labelle and Swaine 2002).

Prior to completing the vignettes, each doctor was asked to complete a short survey. We were interested in individual characteristics— age, gender, level of education, salary, length time in practice, and size of patient list.

Vignette administration and data collection were carried out over approximately one week in each country. The vignettes and short surveys were administered in the same way in each country using

a standardized technique and manual. We scheduled appointments for vignettes to be administered to groups of physicians on site at the hospital or clinic. Vignettes were administered via paper and pencil and took 15–20 minutes to complete per case. We grouped off-duty physicians together in conference rooms or offices to complete the vignettes. Physicians on active duty completed vignettes in between patient visits in examination rooms, emergency rooms, and wards. Sessions for specialists from the tertiary care hospital often took place in large groups ranging from five to 24 physicians at any one time. District hospital physician groups ranged from one to eight physicians and outpatient physician groups ranged in size from one to three physicians. To preserve data integrity, doctors were not told the nature of the condition or disease on which they were being evaluated before completing the vignette.

Analysis

We first did univariate summaries of the data. Next, bivariate analysis was done to allow us to more clearly illustrate the relationships between quality-of-care scores, level of care across country, physician specialization, and study site (public or private). To account for the different means between cases, the quality scores were normalized for the model. The income variable reported in Table 6 used foreign currency conversions to the value of the US Dollar reported on July 31, 2003, when the data were collected.

We used an ANOVA model to estimate the associations between overall scores by case, physician characteristics, study site, and country. Multivariate regression models were developed to evaluate the effect of doctor characteristics on quality score. A separate model was developed to evaluate the associations between income and quality. Interaction terms, for example between specialists and facility, were not significant and therefore dropped from further analysis. The models were tested for clustering of scores by site among the multiple vignettes completed by each physician. The models, however, were robust and the unadjusted model results are presented in

the findings. The independent variables for the quality model were attributed to either the doctor or the care system and were modeled as follows:

$$Quality_i = \alpha + \beta age_i + \sum_{j=1}^{J} \gamma_j P_{ji} + \sum_{k=1}^{K} \omega_k F_{ki} + \sum_{l=1}^{L-1} \theta_l C_{li} + u_i$$

where $Quality_i$ is a quality measure for the ith physician, P_{ji} is a measure of the jth characteristic of the ith physician, F_{ki} measures the kth characteristic of the ith physician's facility, C_{li} is a dummy variable for the lth country, and u_i is a random disturbance term. No missing values were imputed. All statistical evaluation was done using Stata® 8.0.

FINDINGS

The average quality of care score for the 480 cases completed by the 300 doctors in the study was 61.0% (see Figure 1). The scores ranged from 30% to 93%.

Vignette scores showed some modest variation by average score *among* countries, ranging from 60.2% to 62.6% but their differences were not statistically significant. The range of scores *within* each country, however, was broad ($p>0.05$, see Figure 2). The difference between the bottom 5% and the top 5% was 43% in the Philippines (the least variation) to 51% in China (the greatest). This wide variation in quality of care was consistent across facility type (see Figure 3) and by condition (see Figure 4). Analysis of the domain of care again showed wide variation, with variation being the greatest for Testing and Treatment (Figure 5).

We found that physician characteristics were similar among countries, although there was a greater proportion of women physicians and specialists in China. In general, doctors tended to be younger than 44 years old; to have practiced for 8–17 years; and to have been established in the same location 7–8 years. However, doctors in Mexico appeared more likely to be following large panels of patients and to have the highest monthly income (see Table 4).

We then modeled physician characteristics to determine which ones were associated with higher quality of care scores. We included characteristics such as country, ages, level of care, and specialty as well as the patient load of the physician in the model (Table 5). Overall, country of

practice did not explain differences in quality scores. By contrast, younger doctors and women doctors had a significantly higher quality of care than older or male doctors. Type of facility also predicted quality across the five countries. The highest scores tended to be in tertiary testing facilities, followed by private clinics, then public clinics, and finally district facilities. Specialists had higher quality scores than primary care physicians (Table 5). Physicians who followed only a few patients tended to score less than busier doctors but this was not statistically significant. Finally, we ran the same model with physician income and found that income did not predict quality ($p<0.05$).

We then modeled the same physician characteristics against reported monthly income (See Table 6). After controlling for country, the most important predictors of income were level of care, specialty, and age-- physicians less than 35 consistently earned less per month than older doctors. This was in contrast to the previous model used to predict quality, in which older physicians tended to have lower scores than their younger counterparts. Physicians in tertiary care facilities made more than district or private level doctors, who in turn made significantly more than public clinicians. Physicians who reported they cared for less than one thousand patients made less than doctors who had a higher patient load. Of particular note, quality of care did not predict income ($p>0.05$).

When we compared skills among different domains by level of care, we found that physicians at tertiary hospitals typically took better histories, did more complete exams, made diagnoses more accurately, and prescribed the correct treatment more often. These differences were all statistically significant, except for diagnosis. As before, we found wide variation in the performance means by level of care (see Table 7).

Although variation among countries was small, we did find that individual countries appeared to excel in different domains of skills. Mexican physicians appeared to be better at taking histories than those from the other countries, whereas Indian doctors did better examinations. Testing was more accurate in China and diagnostic accuracy was highest in El Salvador. The

Philippines had the highest scores for prescribing treatment. The physical exam, testing, and treatment scores were statistically significant (see Table 8).

DISCUSSION

We conducted a large cross-sectional study of quality of care in five countries. We surveyed 300 doctors who completed 480 identical clinical vignettes on diarrhea, prenatal care, and tuberculosis. We measured quality of care by using physicians' scores on vignettes—validated instruments that show actual quality of care. We found that the overall quality of care was low in all countries and that there was no difference in average quality among countries. The variation in quality of care within countries, however, was quite large, with scores ranging from 30% to 93% of all criteria done correctly by physicians. That variation persisted across facility type, regardless of which clinical condition was measured. The country where care was provided did not predict quality of care. Certain physician characteristics did predict quality. Specifically, younger physicians, specialists, women, and those practicing outside of district hospitals were found to provide a higher level of quality. Interestingly, patient load as measured by listed patients under a doctor's care did not predict quality.

When we analyzed the quality of care by skill domains we found that, overall, physicians scored better in the history and physical exam domains and showed less variation than in the testing or treatment domains. There was also a tendency for physicians in certain countries to perform better in particular domains. Overall, however, the highest standards of care by domain were found in tertiary care hospitals, where doctors were more likely to perform the history and physical correctly and provide the highest quality of treatment.

We also evaluated whether quality of care predicted the income of physicians. After controlling for country-level variables, we found that specialists and doctors working in tertiary, district and private clinics tended to be reimbursed more than non-specialists and those in public facilities. However, higher quality of care was not associated with higher income. Interestingly, one of the greatest predictors of income was the age of the physician. Income and workload

segregated into those that followed more than 1000 patients made less compared to those that did not. The doctor's gender did not affect income although there is the suggestion that at a higher statistical threshold this would be the case (i.e., p>.10).

For us, the most striking finding of the study was the extraordinary variation in quality found within all countries. This has two important implications. First, some physicians in developing countries perform exceptionally well. Thus, insufficient resources are not the sole or even an important predictor of physician capability. Second, poor quality can be addressed by directing remediation toward the poor performers. Strategies targeting poor performers would markedly improve the average quality provided for a given population. And while some of these strategies can target the type of facility or even physicians above a certain age, the overwhelming implication of this study is that quality of care must be measured before physicians can be targeted.

It is interesting that income is not a proxy for quality of care in this study. Indeed, income was best predicted by physician age and specialty. Women, who provided higher quality of care tended to make less than men. Income was also higher in tertiary care hospitals, perhaps reflecting reimbursement policies, patient demand or supply-side effects**.**

This study has several limitations. First, the sample frame was not designed to be representative. Instead, we used a reproducible frame that began with the most well-known tertiary care hospital and then looked at hospitals that referred to the tertiary care hospital and nearby private facilities. Second, physician income was reported by only 245 of the participants; this self-selection may be biased by cultural or other economic factors. Third, we evaluated the quality of care for only three clinical conditions and did so in the outpatient setting.

A major strength of this study was the use of vignettes, which provide a case-mix-adjusted method for measuring quality of care cross-nationally. Vignettes have been used increasingly in developing-country settings to evaluate the quality of care. For example, clinical vignettes were used to evaluate the quality of outpatient clinical practice in Macedonia in comparison with care in the United States (Peabody, Tozija, et al. 2004). Surprisingly, however, the upper 5% of physicians

from the best site in Macedonia performed as well as the upper 25% of physicians in the United States, a finding that supports further investigation and investment in the process elements of care over the structural components. A second strength of the present study is its large sample size and the concurrent collection of the data. A third strength is the focus on the process of care, which is the proximal determinant of health care outcome. Finally, the criteria for measuring process in this study were evidence-based and used international standards of clinical practice.

Implicit in the health care system is that the delivery of high-quality health care services will improve the health of the population (Peabody, Rahman, et al 1999). However, this basic belief has been extraordinarily difficult to prove. One of the most important technological challenges has been how to measure the actual quality of care received by members of a population. This paper introduces the use of vignettes as a way to make cross-national comparisons among several countries. These comparisons did not show that quality was substantially different among countries. However, the enormous variation in quality within countries appears to be a global problem—not one just confined to poor countries. If overall quality is to be improved and if health care is to improve the health status of the general population, it is obvious to us that quality must first be measured (Schuster, et al 2001). Quality scores such as those described here can be reported regularly and given to policy makers, who in turn can design interventions to improve quality. For example, pay-for-performance strategies are gaining attention in some parts of the world; reorganizing systems of care in constant feedback cycles has been implemented in others (Berwick 1998).

Such interventions, intended to rapidly improve clinical practice in a developing-country setting, have already been shown to be successful. In one study, information sessions for private providers on standard case management guidelines for acute respiratory infection, diarrhea, and fever resulted in significant improvements in history-taking, examination, and treatment practices (Chakraborty, D'Souza et al. 2000). In another study, intervention activities—including the provision of basic essential drugs and supplies for the treatment of common childhood diseases, health education, training, and supervision of community health workers, and perhaps most

importantly, changes in clinical practice—were associated with reduced under-five mortality from 155.6 per 1,000 to 61.2 per 1,000 and infant mortality from 114.6 per 1,000 to 40.8 per 1,000 (Afari, Appawu et al. 1995). Techniques similar to those used in this study may be useful in future research to clarify the effectiveness of various intervention-based approaches for improving quality of care. This study suggests that reducing variation in the quality of care is one direction policy makers in developing countries can take.

REFERENCES

Abejo, S. (1987). "Relationship of Infant Mortality and Community Development." *Philippine Population Journal*. 3(1-4): 62-79.

Afari, E., M. Appawu, et al. (1995). "Malaria Infection, Morbidity and Transmission in Two Ecological Zones Southern Ghana." *African Journal of Health Sciences* 2(2): 312-315.

Ashton, C., D. Kuykendall, et al. (1995). "The Association Between the Quality of Inpatient Care and Early Readmission." *Annals of Internal Medicine* 122: 415-421.

Badger, L., F. deGruy, et al. (1995). "Stability of Standardized Patients' Performance in a Study of Clinical Decision Making." *Family Medicine* 27(2): 126-31.

Barker, M. (1983). "Reduction in Perinatal Mortality in Developing Countries: Experiences for Zululand." *Journal of Tropical Pediatrics* 29(5): 268-270.

Beracochea E, D. R., Freeman P, Thomason J. (1995). "Case Management Quality Assessment in Rural Areas of Papua New Guinea." *Tropical Doctor* 25(2): 69-74.

Berggren, W. L., D. C. Ewbank, et al. (1981). "Reduction of Mortality in Rural Haiti Through a Primary Health Care Program." *New England Journal of Medicine* 304(22): 1324-30.

Berwick, D. (1998). "Developing and Testing Changes in Delivery of Care." *Annals of Internal Medicine* 128(8): 651-6.

Blondel, B. (1985) "Some Characteristics of Antenatal Care in 13 European Countries." *British Journal of Obstetrics and Gynaecology* Jun; 92(6):565-8.

Boyce, T. (1991). Self-Selection, Prenatal Care, and Birthweight among Blacks, Whites, and Hispanics in New York City. Cambridge, Mass, National Bureau of Economic Research: 1-18.

Brown, S. S., Ed. (1988). Prenatal Care: Reaching Mothers, Reaching Infants. Committee to Study Outreach for Prenatal Care, Division of Health Promotion and Disease Prevention. Washington DC, National Academy Press.

Bryce, J., M. J. Toole, et al. (1992). "Assessing the Quality of Facility-Based Child Survival Services." *Health Policy and Planning* 7(2): 155-63.

Carney, P. and D. Ward (1998). "Using Unannounced Standardized Patients to Assess the HIV Preventive Practices Of Family Nurse Practitioners and Family Physicians." *The Nurse Practitioner*. 23(2): 56-8, 63, 67-8 passim.

Chakraborty, S., S. A. D'Souza, et al. (2000). "Improving Private Practitioner Care of Sick Children: Testing New Approaches in Rural Bihar." *Health Policy and Planning* 15(4): 400-7.

Collins, T., J. Clark, Et Al. (2002). "Racial differences in How Patients Perceive Physician Communication Regarding Cardiac Testing." *Medical Care* 4(1 suppl): I27-34.

Colliver, J. and M. Swartz (1997). "Assessing Clinical Performance With Standardized Patients." *Journal of the American Medical Association* 278(9): 790-1.

Colliver, J., N. Vu, et al. (1993). "Effects Of Examinee Gender, Standardized-Patient Gender, and Their Interaction on Standardized Patients' Ratings of Examinees' Interpersonal and Communication Skills." *Academic Medicine* 68(2): 153-7.

De Champlain, A., M. Margolis, et al. (1997). "Standardized Patients' Accuracy in Recording Examinees' Behaviors Using Checklists." *Academic Medicine* 72(Supplement 1): S85-S87.

DeGeyndt, W. (1991). Managing Health Expenditures Under National Health Insurance: The Case of Korea. Washington DC, World Bank.

DeGeyndt, W. (1995) "Managing the Quality of Health Care in Developing Countries." WB Technical Paper No. 258. The World Bank. Washington, D.C.

Diette, G. B., E. A. Skinner, et al. (2001). "Comparison of Quality of Care by Specialist and Generalist Physicians as Usual Source of Asthma Care for Children." *Pediatrics* 108(2): 432-7.

Donabedian, A. (1980). *The Definition of Quality and Approaches to Its Assessment*. Ann Arbor, MI, Health Administration Press.

Donabedian, A. (1988). "The Quality of Care: How Can It be Assessed?" *Journal of the American Medical Association* 260(12): 1743-8.

Donohoe, MT. (1998). "Comparing Generalist and Specialty Care: Discrepancies, Deficiencies, and Excesses." *Archives of Internal Medicine* Aug 10-24; 158(15):1596-608.

Dresselhaus, T. R., J. Luck, et al. (2002). "The Ethical Problem of False Positives: A Comparison of Standardized Patients and the Medical Record." *Journal of Medical Ethics* 28: 291-294.

Dresselhaus, T. R., J. W. Peabody, et al. (2000). "Measuring Compliance with Preventive Care Guidelines: Standardized Patients, Clinical Vignettes, and the Medical Record." *Journal of General Internal Medicine* 15(11): 782-8.

Druss, B., W. Bradford, et al. (2001). "Quality of Medical Care and Excess Mortality in Older Patients with Mental Disorders." *Archives of General Psychiatry* 58(6): 565-72.

Dumont, A., L. De Bernis, et al. (2002). "Maternal Morbidity and Qualification of Health-Care Workers: Comparison Between Two Different Populations in Senegal." *Journal de Gynecologie, Obstetrique et Biologie de la Reproduction* 31(1): 70-9.

Fihn, SD. (2000) "The Quest to Quantify Quality." *Journal of the American Medical Association* April 5; 283(13):1740-2

Glassman, P. A., J. Luck, et al. (2000). "Using Standardized Patients to Measure Quality: Evidence from the Literature and a Prospective Study." *The Joint Commission Journal on Quality Improvement* 26(11): 644-53.

Goldman, R. (1992). "The Reliability of Peer Assessments of Quality of Care." *Journal of the American Medical Association* 267: 958-960.

Goulding, M. (2004). "Inappropriate Medication Prescribing for Elderly Ambulatory Care Patients." *Archives of Internal Medicine* 164(3): 305-12.

Harrold, L. R., T. S. Field, et al. (1999). "Knowledge, Patterns of Care, and Outcomes of Care for Generalists and Specialists." *Journal of General Internal Medicine* 14(8): 499-511.

Institute of Medicine (IOM). (1993) *Access to Health Care in America*. Washington, D.C.: National Academy Press

Institute of Medicine (IOM). (2001) *Crossing the Quality Chasm*. Washington, D.C.: National Academy Press

Jans, M., F. Schellevis, et al. (2000). "Improving General Practice Care of Patients with Asthma or Chronic Obstructive Pulmonary Disease: Evaluation of a Quality System." *Effective Clinical Practice* 3(1): 16-24.

Jans, M. P., F. G. Schellevis, et al. (2001). "Health Outcomes of Asthma and COPD Patients: The Evaluation of a Project to Implement Guidelines in General Practice." *International Journal for Quality in Health Care* 13(1): 17-25.

Kerr, E., R. Gerzoff, et al. (2004). "Diabetes Care Quality in the Veterans Affairs Health Care System and Commercial Managed Care: the TRIAD Study." *Annals of Internal Medicine* 141(4): 272-81.

Kravitz, R., S. Greenfield, et al. (1992). "Differences in the Mix of Patients Among Medical Specialties and Systems of Care: Results from the Medical Outcomes Study." *Journal of the American Medical Association* 267: 1617-1623.

Krieger, J. (1989). The Quality of Prenatal Care Received by Medicaid Beneficiaries Enrolled in Managed Care Plans in the state of Washington. Thesis (M.P.H.)--University of Washington.

Labelle, J. and B. Swaine (2002). "Reliability Associated with the Abstraction of Data from Medical Records for Inclusion in an Information System for Persons with a Traumatic Brain Injury." *Brain Injury* 16(8): 713-27.

Lavy, V., J. Strauss, et al. (1996). "Quality of Health Care, Survival and Health Outcomes in Ghana." Journal of Health Economics 15(3): 333-57.

Loevinsohn, B. P., E. T. Guerrero, et al. (1995). " Improving Primary Health Care Through Systematic Supervision: A Controlled Field Trial." *Health Policy and Planning* 10(2): 144-153.

Luck, J., J. Peabody, et al. (2000). "How Well Does Chart Abstraction Measure Quality? A Prospective Comparison of Standardized Patients with the Medical Record?" *American Journal of Medicine* 108(8): 642-9.

Luck, J. and J. Peabody (2002). "Using Standardised Patients to Measure Physicians' Practice: Validation Study Using Audio Recordings." *British Medical Journal* 325(678).

McCormick, M. (1985). "The Contribution of Low Birth Weight to Infant Mortality and Childhood Morbidity." *New England Journal of Medicine* 312(2): 82-90.

McDonald, C., J. Overhage, et al. (1997). "A Framework for Capturing Clinical Data Sets from Computerized Sources." *Annals of Internal Medicine* 127(8 Pt 2): 675-82.

McGlynn, E. A., S. M. Asch, et al. (2003). "The Quality of Health Care Delivered to Adults in the United States." *New England Journal of Medicine* 348(26): 2635.

Nagaya, K., M. Fetters, et al. (2002). "Causes of Maternal Mortality in Japan." *Journal of the American Medical Association* 283(20): 2661-7.

Nolan, T., P. Angos, et al. (2001). "Quality of Hospital Care for Seriously Ill Children in Less-Developed Countries." *Lancet* 357(9250): 106-10.

Norman, G., D. Davis, et al. (1993). "Competency Assessment of Primary Care Physicians as Part of a *peer review program.*" *Journal of the American Medical Association* 270: 1046-1051.

O'Connor, G., H. Quinton, et al. (1999). "Geographic Variation in the Treatment of Acute Myocardial Infarction: The Cooperative Cardiovascular Project." *Journal of the American Medical Association* 281(7): 627-33.

Palmer, R., T. Louis, et al. (1985). "A Randomized Controlled Trial of Quality Assurance in Sixteen Ambulatory Care Practices." *Medical Care* 23: 751-770.

Peabody, J., P. Gertler, et al. (1998). "The Policy Implications of Better Structure and Process on Birth Outcomes in Jamaica." *Health Policy* 43(1): 1-13.

Peabody, J. and J. Luck (1998). "How Far Down the Managed Care Road? A Comparison of Primary Care Outpatient Services in a Veterans Affairs Medical Center and a Capitated Multispecialty Group Practice." *Archives of Internal Medicine* 158: 2291-2299.

Peabody, J., J. Luck, et al. (2000). "Comparison of Vignettes, Standardized Patients, and Chart Abstraction: A Prospective Validation Study of 3 Methods for Measuring Quality." *Journal of the American Medical Association* 283(13): 1715-22.

Peabody, J., J. Luck, et al. (2004 *a*) . "Assessing the Accuracy of Administrative Data in Health Information Systems." *Medical Care* Nov; 42(11):1066-72.

Peabody, J., J. Luck, et al. (2004 *b*). "Measuring the Quality of Physician Practice by Using Clinical Vignettes: A Prospective Validation Study." *Annals of Internal Medicine* Nov 16; 141(10):771-80.

Peabody, J., M. Rahman, et al. (1999). *Policy and Health: Implications for Development in Asia.* Cambridge, UK, Cambridge University Press.

Peabody, J., O. Rahman, et al. (1994). "Quality of Care in Public and Private Primary Health Care Facilities: Structural Comparisons in Jamaica." *Bulletin of the Pan American Health Organization* 28(2): 122-141.

Peabody, J., F. Tozija, et al. (2004). "Using Vignettes to Compare the Quality of Care Variation in Economically Divergent Countries." *Health Services Research* Dec; 39(6 Pt 2):1951-70

Peabody, J. W. (1995). "Will Measuring the Quality of Antenatal Care Tell Us How It Works?" RGSD-112 Santa Monica, CA., RAND.

Rethans, J. and C. van Boven (1987). "Simulated Patients in General Practice: A Different Look at the Consultation." *British Medical Journal* (Clin Res Ed) 294(6575): 809-12.

Rosen, A., A. Ash, et al. (1995). "The Importance of Severity of Illness Adjustment in Predicting Adverse Outcomes in the Medicare Population." *Journal of Clinical Epidemiology* 48: 631-643.

Salem-Schatz, S., G. Moore, et al. (1994). "The Case for Case-Mix Adjustment in Practice Profiling: When Good Apples Look Bad." *Journal of the American Medical Association* 272: 871-874.

Salmon, J., J. Heavens, et al. (2003). *The Impact of Accreditation on the Quality of Hospital Care: KwaZulu-Natal Province, Republic of South Africa*. Bethesda, MD, Quality Assurance Project, University Research Co., LLC.

Schuster, M., E. McGlynn, et al. (1998). "How Good Is the Quality of Health Care in the United States?" *Milbank Quarterly* 76(4): 517-63.

Schuster, M. A., E. A. McGlynn, C. B. Pham, M. D. Spar, and R. H. Brook. (2001). "The Quality of Health Care in the United States: A Review of Articles Since 1987." In: Medicine I. O., eds., *Crossing the Quality Chasm*. Washington, D.C.: National Academy Press**:** 243-321

Seddon, M. E., J. Z. Ayanian, et al. (2001). "Quality of Ambulatory Care After Myocardial Infarction Among Medicare Patients by Type of Insurance and Region." *American Journal of Medicine* 11(1): 24-32.

Silverman, M., M. A. Terry, et al. (2002). "The Role of Qualitative Methods in Investigating Barriers to Adult Immunization." *Qualitative Health Research* 12(8): 1058-1075.

Skinner, J., E. Fisher, et al. (2001). "Efficiency of Medicare," NBER Working Paper 8395. Cambridge, Mass, National Bureau of Economic Research.

Swartz, M. and J. Colliver (1996). "Using Standardized Patients for Assessing Clinical Performance: an Overview." *Mt Sinai Journal Of Medicine* 63(3-4): 241-9.

A Vision for India's Health System, Conference, November 15-16, 2001 New Delhi, India.

Walker, G. J., D. E. Ashley, et al. (1988). "The Quality of Care Is Related to Death Rates: Hospital Inpatient Management of Infants with Acute Gastroenteritis in Jamaica." *American Journal of Public Health* 78(2): 149-152.

Weinberg, M. (2001). "Reducing Infections Among Women Undergoing Cesarean Section in Colombia by Means of Continuous Quality Improvement Methods." *Archives of Internal Medicine* 161(19): 2357-2365.

Wennberg, J. E., E. S. Fisher, et al. (2002). "Geography and the Debate Over Medicare Reform." *Health Affairs (Millwood)*. Supp Web Exclusives: W96-114.

WHO (2000). *The World Health Report 2000 - Health Systems: Improving Performance.* Geneva, World Health Organization.

WHO (2003). *The World Health Report 2003, Shaping the Future.* Geneva, World Health Organization.

World Bank Group, T. (2004). *Tuberculosis Control Project.* Washington DC, World Bank.

Wu, L. and C. Ashton (1997). "Chart Review: A Need for Reappraisal." *Evaluation & the Health Professions*. 20: 146-163.

Zadnik, K., M. Mannis, et al. (1998). "Inter-Clinician Agreement on Clinical Data Abstracted from Patients' Medical Charts." *Optometry and Vision Science* 75(11): 813-6.

Table 1. Characteristics of Different Quality Measurement Methods

Method	Cost	Bias	Data Collection	Case-Mix Variation	Ethical Issues	Advantages	Disadvantages
Chart abstraction	High due to need for expert abstractor (Norman, Davis et al. 1993; Ashton, Kuykendall et al. 1995)	Subject to recording bias (because of time con-straints on outpatient visits) (Ashton, Kuykendall et al. 1995)	Must be done by trained personnel (Norman, Davis, et al 1993). No set of established standards to use for abstraction of charts.	Insufficient adjustment for case-mix variation, limiting direction comparisons of quality of care across different sites or delivery systems (Rosen, Ash et al. 1995)	Can be safeguarded with appropriate steps	Available after every patient contact	Primarily validated in inpatient settings (Wu and Ashton 1997). Recording practices may vary from setting to setting and country to country. Does not accurately record the provider-patient interaction
Direct observation and recorded visits	Potentially high	Hawthorne effect exists if physicians perform better under observation (Luck and Peabody 2002); variation between observers inevitable	Requires trained expert.	None available	Informed consents required: blinding both patient and provider would be unethical		No published standards. Patient and physician may be uncom-fortable being observed.
Administra-tive data	Low	Multiple: recording, payment, and system-based factors influence data base	Accuracy and reliability of data depend on who entered data; not subject to strict regulations.	No case-mix adjustment possible. Strong tendency toward under-reporting of secondary diagnoses	Can be minimal	Generally available; becoming more and more available in developing countries	Often compromised by data entry and recording inaccuracies. Notes on history- taking or physical examination not generally included.

Standardized patients (SPs)	Very high. SPs are often well-paid, trained actors.	Minimal	Extremely complex	By design, completely case-mix adjusted	No patients involved	Unannounced SPs provide especially accurate measures of clinical practice. Can be used to compare quality between different sites.	Incurs opportunity cost for the physician. Daunting logistics requirements. Useful only for adult conditions that can be simulated
Vignettes	Low	Must be constructed in a way that eliminates effort-related bias	Convenient	By design, completely case-mix adjusted	No patients involved	Responsive to variation. Readily accepted by providers; refusal rates are typically low, providing a more accurate statistical measure of quality. Validated as a measure of actual clinical practice in recent studies	Limited to an evaluation of a single clinical visit. Critics argue that vignettes are not the same as actual clinical practice.

Table 2. Tertiary Hospitals Used as Reference Facilities

	Tertiary Hospital
China	Peking Union Medical College
El Salvador	Instituto Nacional de la Seguridad Social
India	All India Institute of Medical Sciences
Mexico	Instituto Nacional de Ciencias Medicas y Nutricion
Philippines	Philippine General Hospital

Table 3. Selected Health Indicators for Study Countries

	China	El Salvador	India	Mexico	Philippines
Population (1,000s) 2002*	1,302,307	6,415	1,049,549	101,965	78,580
Life expectancy—years (male/female) 2002*	69.6/72.7	66.5/72.8	60.1/62.0	71.7/77.0	65.1/71.7
Probability of dying under age 5 (male/female) per 1000, 2002*	31/41	36/34	87/95	30/24	39/33
% of children under 5 with diarrhea in the two weeks prior to survey **	N/A	19.8% (1998)	19.2% (1999)	9.7% (1993)	7.4% (1998)
Maternal mortality rate (per 100,000 live births), 2002***	272	83	344	44	540
TB Prevalence (per 100,000), 2002***	272	83	344	44	540
DOTS Detection Rate, 2002***	27%	57%	31%	73%	58%

* WHO, 2003, World Development Report 2003.
** UNICEF Database on Diarrhoeal disease and ORT. Note: seasonal variation and inconsistent timing of surveys make country data incomparable. http://www.childinfo.org/eddb/Diarrhoea/database.htm.
*** United Nations Statistics Division, http://millenniumindicators.un.org/.

Table 4. Physician Characteristics by Country

Country

Characteristics of 300 Physicians	China	El Salvador	India	Mexico	Philippines	Total Number
Level of Care in Sample						
Tertiary	24	25	24	24	26	123
District	8	24	9	24	8	73
Public clinic	24	14	9	8	8	63
Private clinic	8	9	8	8	8	41
Total	64	72	50	64	50	300
Age Group						
Under 35	32	38	24	15	21	130
35–44	15	18	13	24	19	89
45–54	13	13	12	18	7	63
Over 55	4	3	1	7	3	18
Total	64	72	50	64	50	300
Gender						
Male	15	38	30	41	19	143
Female	49	34	20	22	31	155
Total	64	72	50	64	50	300
Specialty						
Primary care	16	25	26	23	28	118
Internal medical	17	18	7	17	7	66
OB/GYN	23	15	9	14	8	69
Pediatrics	8	14	8	10	7	47
Total	64	72	50	64	50	300
Percentage of Physicians Under 35						
Tertiary care facilities	67	48	71	17	35	47
District care facilities	25	83	56	25	63	52
Public clinic care facilities	58	29	0	13	13	32
Private clinic care facilities	0	22	25	63	75	37
Total percentage of physicians under 35	50	53	48	25	42	44
Monthly Income (U.S. Dollars in 2003)						
Mean	261	155	535	1380	858	
Standard deviation	134	146	221	772	1296	
Mean of top 25 percent	437	336	828	2477	2218	
Years Practicing Medicine						
Mean	15	9	13	15	12	
Standard deviation	11.1	7.5	8.8	10.9	9.5	
Years in the Same Location						
Mean	8	7	7	11	8	
Standard deviation	6.5	7.1	7.5	8.2	6.6	
Percentage of Physicians Following:						
0–500 listed patients	78	44	44	11	56	
501–1000 listed patients	8	14	10	5	14	
1001–1500 listed patients	11	6	2	5	4	
1501–2500 listed patients	0	8	8	5	4	
Over 2500 *Listed Patients*	0	24	24	31	22	
Not Answered						63
Total Number of Physicians = 300						

Table 5. Model of Quality by Country and Other Characteristics

Dependent Variable: Quality of Care (Z-value of Overall Quality Score)

Independent Variables	Coefficient	Standard Error	P-value
Country Variables			
China	-0.16	0.15	0.28
El Salvador	-0.15	0.14	0.29
India	0.11	0.14	0.43
Mexico	0.13	0.16	0.39
Philippines	(dropped)		
Age and Gender			
Age less than 35	0.39	0.19	0.04
Age among 35 to 44	0.26	0.19	0.18
Age among 45 to 54	0.22	0.20	0.26
Age over 55	(dropped)		
Gender (Female)	0.27	0.09	0.00
Level of Care			
Tertiary	0.58	0.14	0.00
Public Clinic	0.40	0.13	0.00
Private Clinic	0.45	0.13	0.00
District	(dropped)		
Specialty			
Internal Medical	0.38	0.14	0.01
OB/GYN	0.44	0.15	0.00
Pediatrics	0.41	0.17	0.02
Primary Care Physician	(dropped)		
Number of Patients Followed			
0 to 500 Listed Patients	-0.09	0.17	0.58
501 to 1000 Listed Patients	-0.30	0.20	0.14
1001 to 1500 Listed patients	0.09	0.24	0.71
1501 to 2500 Listed Patients	0.26	0.24	0.29
Over 2500 Listed Patients	0.04	0.16	0.81
Not Answered	(dropped)		
Constant	-0.89	0.27	0.00
Number of Observations	480		
F statistic	4.65		
R-square	0.13		

Table 6. Income and Quality

Dependent Variable: Income Reported (log 2003 US Dollars)

Independent Variables	Coefficient	Standard Error	P-value
Overall Z-score value	0.07	0.04	0.09
Country Variables			
China	-0.75	0.14	0.00
El Salvador	-1.51	0.13	0.00
India	-0.16	0.13	0.24
Philippines	0.36	0.15	0.02
Mexico	(dropped)		
Age			
Age less than 35	-0.55	0.17	0.00
Age among 35 to 44	-0.20	0.17	0.26
Age among 45 to 54	0.05	0.17	0.76
Age over 55	(dropped)		
Gender (Woman)	-0.15	0.08	0.08
Level of Care			
Tertiary	0.60	0.12	0.00
District	0.39	0.12	0.00
Private Clinic	0.35	0.14	0.01
Public Clinic	(dropped)		
Specialty			
Internal Medical	-0.32	0.12	0.01
Pediatrics	-0.45	0.13	0.00
Primary Care Physician	-0.37	0.12	0.00
OB/GYN	(dropped)		
Number of Patients Followed			
0 to 500 Listed Patients	-0.33	0.15	0.03
501 to 1000 Listed Patients	-0.35	0.18	0.05
1001 to 1500 Listed patients	-0.06	0.21	0.77
1501 to 2500 Listed Patients	-0.03	0.21	0.90
Over 2500 Listed Patients	-0.10	0.15	0.49
Not Answered	(dropped)		
Constant	6.82	0.25	0.00
Number of Physicians	245*		
F statistic	27.83		
R-square	0.71		

* 55 of 300 physicians did not provide income information.

Table 7. Mean and Standard Deviation of Domain Score by Level of Care

	History		Physical Exam		Testing		Diagnosis		Treatment	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Tertiary	**0.66**	0.16	**0.75**	0.19	0.56	0.33	**0.81**	0.24	**0.59**	0.23
District	0.58	0.16	0.63	0.21	**0.61**	0.32	0.77	0.27	0.49	0.23
Public clinic	0.63	0.14	0.66	0.22	0.55	0.31	0.75	0.27	0.51	0.20
Private clinic	0.63	0.14	0.68	0.21	0.55	0.35	0.77	0.25	0.52	0.22
	Facility Type with Highest Quality in Each Domain									
Highest Score	**Tertiary**		**Tertiary**		**District**		**Tertiary**		**Tertiary**	
P-value	<0.01		<0.01		0.13		0.11		<0.01	

Table 8. Mean and Standard Deviation of Domain Score by Country

	History		Physical Exam		Testing		Diagnosis		Treatment	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
China	0.625	0.165	0.627	0.253	**0.703**	0.263	0.750	0.305	0.491	0.241
El Salvador	0.612	0.134	0.672	0.198	0.618	0.304	**0.838**	0.212	0.494	0.216
India	0.629	0.152	**0.715**	0.198	0.410	0.332	0.718	0.266	0.544	0.230
Mexico	**0.630**	0.155	0.713	0.201	0.610	0.314	0.785	0.258	0.523	0.222
Philippines	0.625	0.148	0.672	0.208	0.500	0.348	0.788	0.246	**0.572**	0.197
	Country with Highest Domain									
Highest score	**Mexico**		**India**		**China**		**El Salvador**		**Philippines**	
P-value	0.230		0.000		0.000		0.790		0.000	

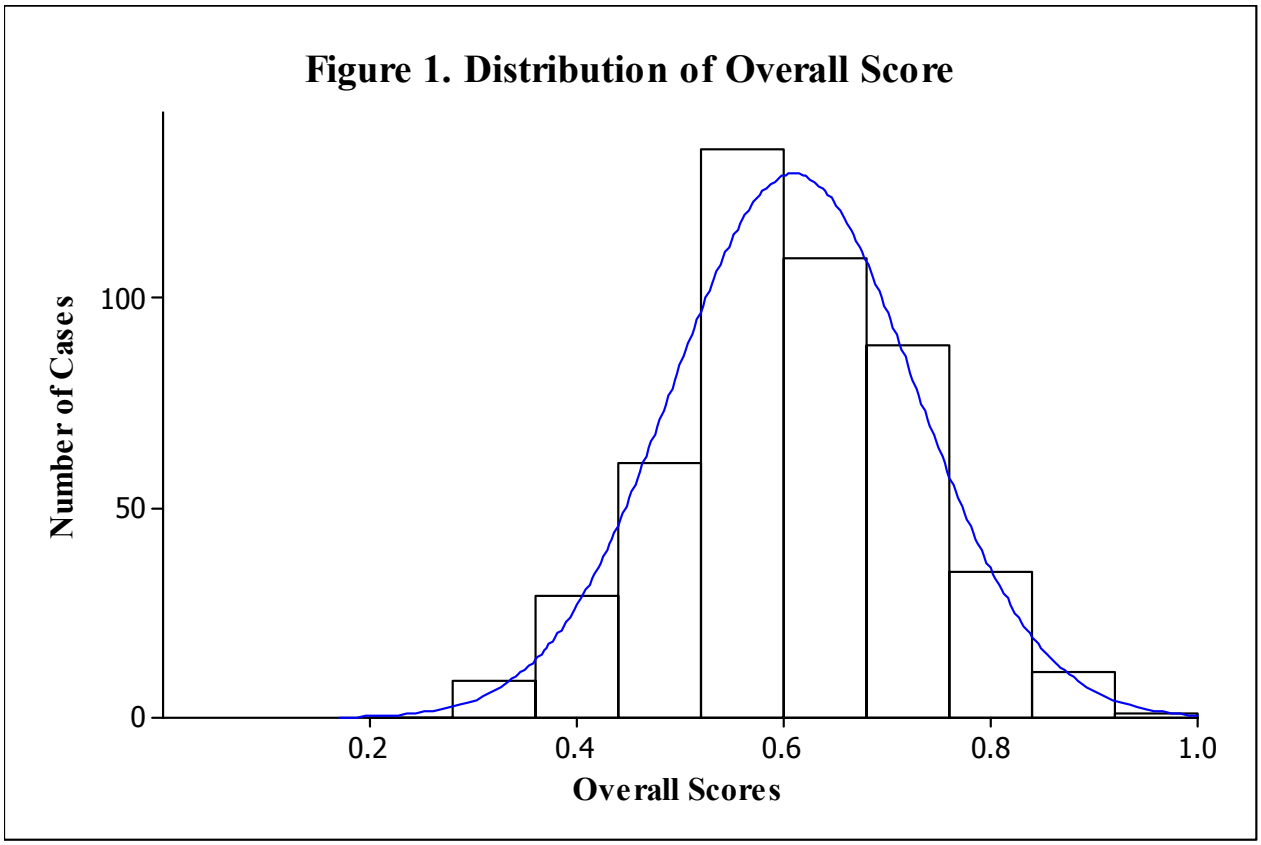


Figure 1. Distribution of Overall Score
Number of Cases
100
50
0
0.2
0.4
0.6
0.8
1.0
Overall Scores

Figure 2. Comparison of Overall Scores Across Countries

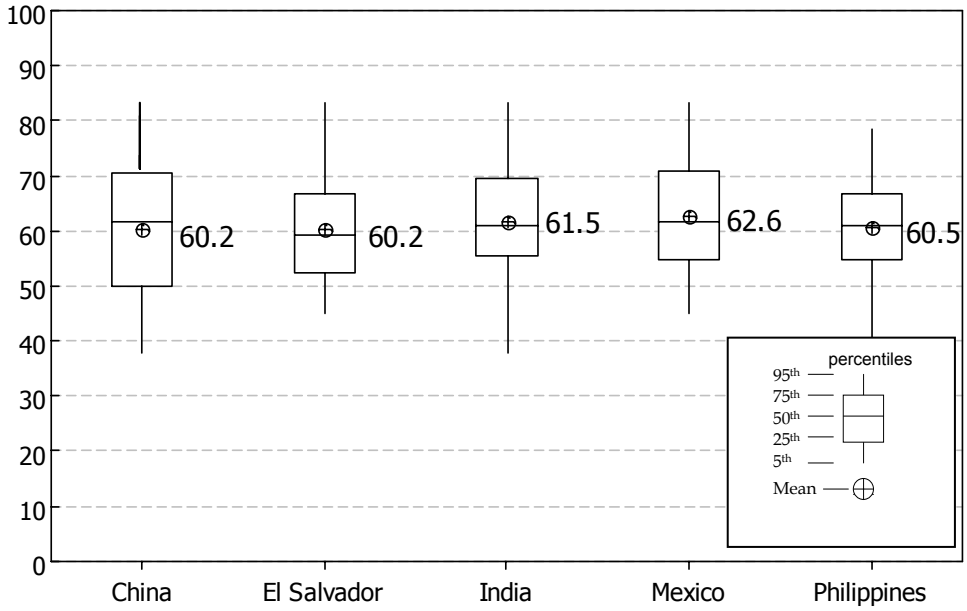


Figure 3. Variation in Overall Scores by Facility Type

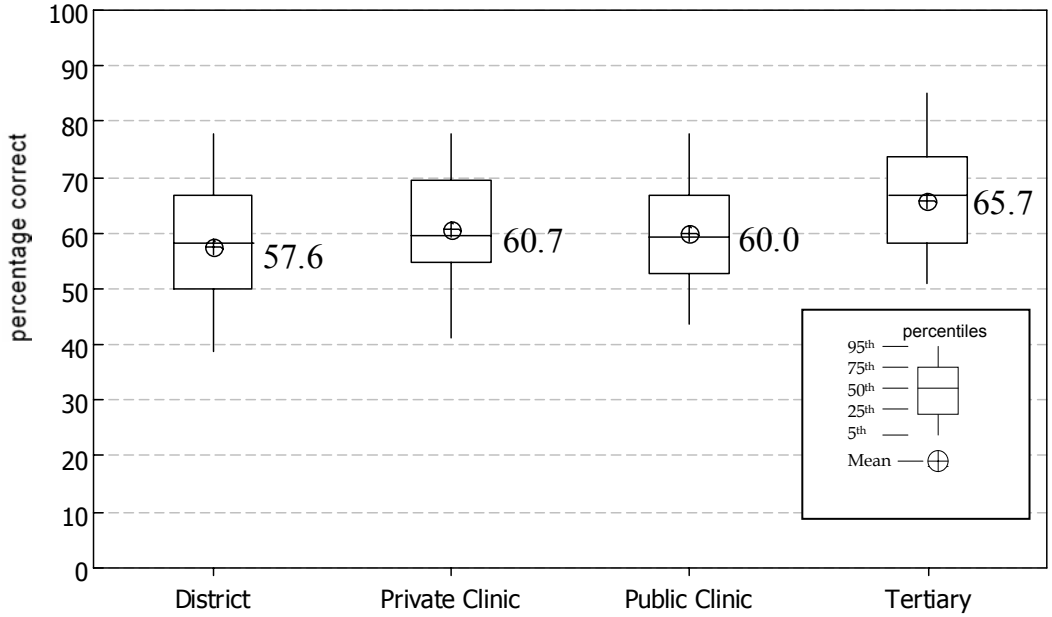
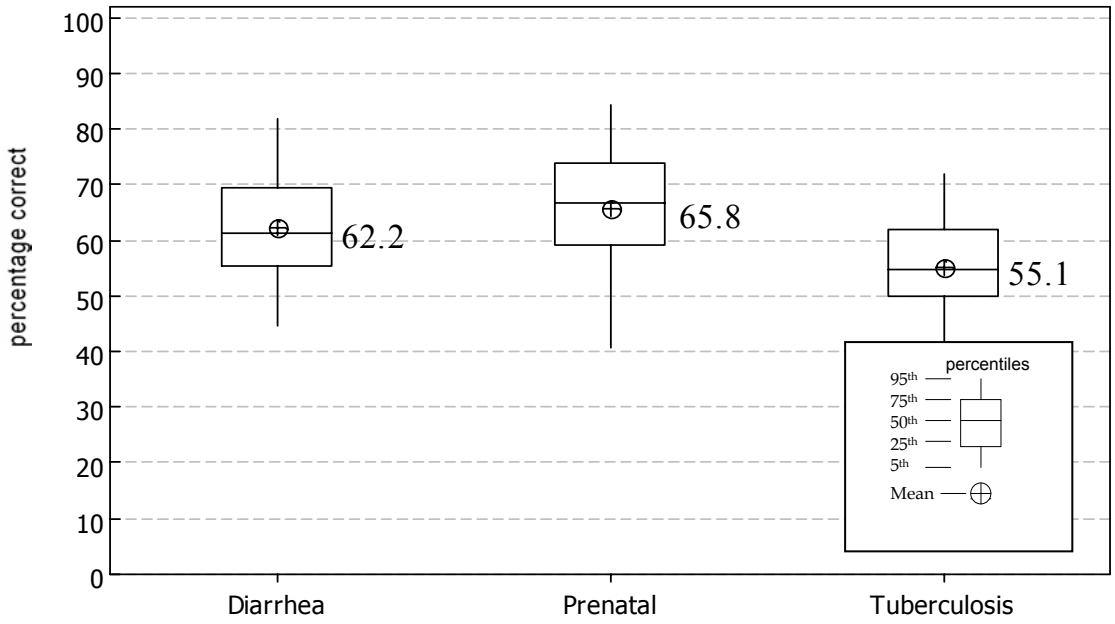
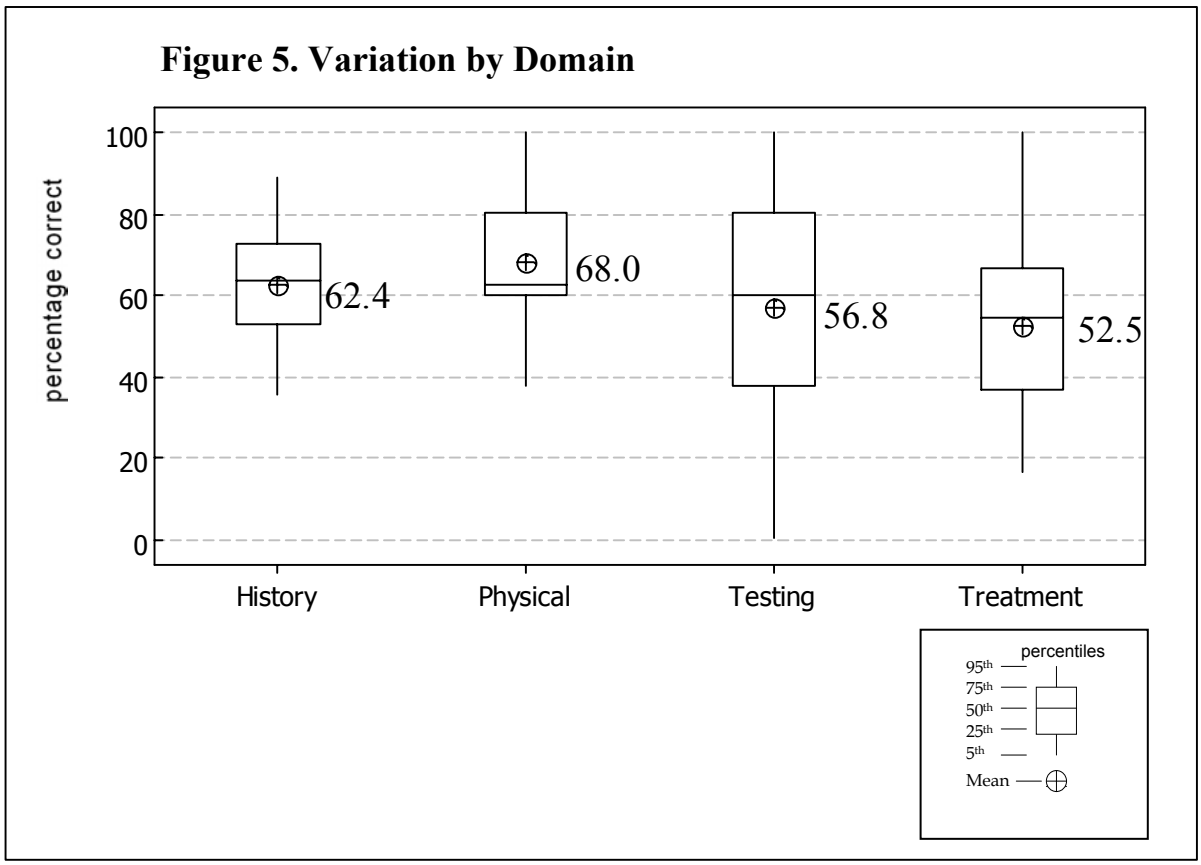


Figure 4. Variation in Overall Scores by Condition





Note: Diagnosis is omitted from this analysis because by construct there were only 2–3 items available for determining a distribution.